applied sciences

MDPI

Article

TREASURE: Text Mining Algorithm Based on Affinity Analysis and Set Intersection to Find the Action of Tuberculosis Drugs against Other Pathogens

Pradeepa Sampath [1], Nithya Shree Sridhar [1], Vimal Shanmuganathan [2] and Yangsun Lee [3,*]

[1] School of Computing, SASTRA Deemed University, Thanjavur 613402, India; pradeepa.pradee@gmail.com (P.S.); nithshree17@gmail.com (N.S.S.)
[2] Department of Computer Science and Engineering, Ramco Institute of Technology, Rajapalayam 626117, India; svimalphd@gmail.com
[3] Division of Computer Engineering, Hanshin University, Osan 18101, Korea
* Correspondence: yslee48@gmail.com

Abstract: Tuberculosis (TB) is one of the top causes of death in the world. Though TB is known as the world's most infectious killer, it can be treated with a combination of TB drugs. Some of these drugs can be active against other infective agents, in addition to TB. We propose a framework called TREASURE (Text mining algoRithm basEd on Affinity analysis and Set intersection to find the action of tUberculosis dRugs against other pathogEns), which particularly focuses on the extraction of various drug–pathogen relationships in eight different TB drugs, namely pyrazinamide, moxifloxacin, ethambutol, isoniazid, rifampicin, linezolid, streptomycin and amikacin. More than 1500 research papers from PubMed are collected for each drug. The data collected for this purpose are first preprocessed, and various relation records are generated for each drug using affinity analysis. These records are then filtered based on the maximum co-occurrence value and set intersection property to obtain the required inferences. The inferences produced by this framework can help the medical researchers in finding cures for other bacterial diseases. Additionally, the analysis presented in this model can be utilized by the medical experts in their disease and drug experiments.

Keywords: tuberculosis drugs; pathogens; PubMed; affinity analysis; set intersection

Citation: Sampath, P.; Sridhar, N.S.; Shanmuganathan, V.; Lee, Y. TREASURE: Text Mining Algorithm Based on Affinity Analysis and Set Intersection to Find the Action of Tuberculosis Drugs against Other Pathogens. *Appl. Sci.* **2021**, *11*, 6834. https://doi.org/10.3390/app11156834

Academic Editor: Federico Divina

Received: 9 June 2021
Accepted: 21 July 2021
Published: 25 July 2021

1. Introduction

According to the World Health Organization (WHO) report, the number of TB cases went from 2.2 million to 2.8 million in the year 2015 [1]. It also stated that the global estimates went up from 9.6 million to 10.4 million, and the number of deaths caused by TB doubled in India. In 2016, the greatest number of new cases of Multidrug-resistant TB (MDR-TB) was reported in India. By WHO estimates in 2017, around 400,000 people died among the 2.7 million people affected by TB [2]. In the year 2019, around 2.64 million TB cases were reported by the WHO in India, and estimates show that 40% of the Indian population was affected by TB bacteria. Many journals, conferences, and patents are dedicated to the study of TB, TB survey and anti-TB drugs in PubMed. PubMed is a free search engine to primarily access the MEDLINE database and it comprises more than 32 million citations and abstracts for biomedical literature [3]. Manyhealth related information such as diseases, their symptoms, prevention, treatment, many plants, their activity, usage, etc. are present in this database.

Tuberculosis (TB) is a contagious and an infectious bacterial disease caused by a bacterium named *Mycobacterium tuberculosis* that mainly affects the lungs [4]. It can also affect other parts of the body like the spine, kidney and brain. TB bacteria are spread through air when an infected person sneezes, coughs, or speaks. It is a curable disease. A combination of antibiotic medications is given for patients with active symptoms and they

must undergo a long course of treatment [5,6]. According to WHO, around 63 million lives were saved through TB treatment since the year 2000.

The demand for TB drugs is always high, and over the years, many anti-TB agents have been developed. The treatment for pulmonary TB (lungs affected) involves two antibiotics for six months and an additional two antibiotics for the first two months. The first-line anti-TB agents such as rifampicin, isoniazid, pyrazinamide and ethambutol are the prescribed antibiotics [7]. The same combination of medications can be used for the treatment of Extrapulmonary TB (outside the lungs). The drugs are usually developed to attack or kill the *bacterium* that causes the disease. Each drug has its own specific molecular target to which it binds and produces its pharmacological effect [8]. Their mechanism of action may also be effective against other pathogens. Thus, these drugs can also be used in the treatment of other diseases. This is known as drug repurposing and it is gaining attention due to the significant rise in the costs of pharmaceutical R&D. The pharmaceutical companies are looking for such repurposing strategies [9,10].

The existing text mining algorithms can be used to classify drugs with respect to a particular disease [11,12]. However, there are no reports of any inferences drawn on the activity of drugs against a spectrum of pathogens, in addition to the pathogen responsible for an infectious disease. This motivated the need to design a technology-based solution by using TB drugs-based PubMed abstracts as the source dataset. For this purpose, we collected 5 years of recent abstracts on various TB drugs research from PubMed. The large volume of extracted data should be mined to find the underlying patterns. The proposed framework, TREASURE, finds various relations among the preprocessed data, with the help of affinity analysis, and it uses set intersection property to filter out these patterns among the relation sets based upon their occurring frequency. This framework is applied on various TB drugs datasets to determine the various other infections these drugs are effective against in addition to TB. This method might help various researchers in the field of drug discovery and drug interaction studies, doctors and also the pharmaceutical companies.

2. Literature Review

The gene–disease relationships have a great significance in diagnosis, treatment and prevention of diseases. Though these associations are deeply investigated by the researchers, much of their underpinnings are yet to be explained. Text mining was performed on documents from the PubMed database to predict the gene–disease relationships based on the cosine similarity between the gene vectors and disease vectors [13]. This method integrates the MeSH database, co-occurrence methods and term weight to predict gene-disease relationships. Chemical and drug information is accumulated in all sorts of text documents like industry reports, patents and scientific articles. This has led to the development of many text mining applications. The PubMed biomedical literature database is a valuable source of information. An R package was developed by combining the advantages of existing text mining algorithms, to analyze the PubMed abstracts [14]. A review on the tools, methods and applications of text mining for chemical compounds, was presented to determine their structures and identify relationships between chemicals and other entities [15].

Latent Dirichlet Allocation is a probabilistic topic model that aims to give a topic perspective solution. Researchers have proposed various topic models based on LDA. An algorithm called Bio-LDA was introduced to identify latent topics using biological terms and uncover putative relations among bio-terms and topics [16]. A survey was presented to discover trends, research development and intellectual structures of topic modelling based on LDA [17]. To extract research topics from Alzheimer's disease-related papers, the LDAP framework was proposed, which combined LDA with an affinity propagation model algorithm [18]. Various analyses, like research trends analysis, were performed on the results. Another model, named latent semantic analysis, was used to semantically interface PubMed abstracts to gene ontology [19]. The PubMed abstracts for this model were obtained based on the semantic similarity between the user query and the abstracts.

Three keyword extraction models were proposed in which the first one was based on LSA and right singular matrix, and the other two models were based on Shannon entropy [20]. The proposed models are not length dominating, and they have a low redundancy. For the analysis of patent data, an intelligent system based on principle component analysis and logistics was proposed [21]. This system extracts the features from the patents database and classifies them into categories such as software, biological, business and chemical.

The prediction of TB survivability has been a challenging problem for many years. A study was presented on various data mining approaches that have been utilized for Tuberculosis diagnosis and prognosis, and a best prediction model for TB survivability was developed accordingly [22]. A regulatory network was proposed to detect the action of genes required for *Mycobacterium tuberculosis* (*M.tb*) persistence, using text mining models [23]. Its purpose was to suggest candidates for new drug targets and to provide fresh insights on the persistence mechanism of *M.tb*. A method to find close association rules within TB data, was proposed and applied on the dataset containing the real medical records of TB patients [24]. This method determined the association of one symptom to another. Different types of media content analysis are carried out for different application.

Taking all these as a motivation, this paper proposes a model to deal with PubMed datasets in the identification of the various pathogens the TB drugs are active against, in addition to *M.tb*. The analyzed TB drugs can be effective against some other diseases, caused by the pathogens that these drugs work against.

3. Materials and Methods

This article proposes a TREASURE framework, that follows affinity analysis to find various relations between the dataset and filters out the patterns among these records via set intersection and occurring frequency. Algorithm 1 gives the outline of the TREASURE model. Thousands of documents pertaining to each drug are collected from the PubMed database. The collected documents are preprocessed as they undergo tokenizing, stemming, stop words removal and tf-idf calculation phases. The preprocessed data are visualized through word cloud to give an idea about the frequently occurring words in the collected dataset. Then, the preprocessed data undergoes an affinity analysis phase, which determines various co-occurring relationships among the words in the dataset and generates the relation records accordingly. These relation records are then filtered, based on maximum co-occurrence value and set intersection property, to obtain the resultant set. Thus, different resultant sets for different TB drugs are obtained and are then analyzed to provide various inferences. The overall architecture of the TREASURE framework is depicted in Figure 1.

Algorithm 1 Outline of TREASURE model

Input: Abstracts from PubMed Database
Output: A filtered resultant set containing frequently occurring word patterns
1. Data preprocessing
   1.1 Tokenizing
   1.2 Remove punctuations and stop words
   1.3 Stemming
   1.4 TF-IDF calculation
2. Generate the relation records via affinity analysis
3. Based on the co-occurrence value of each element and set intersection property, the resultant set is filtered out from the relation records

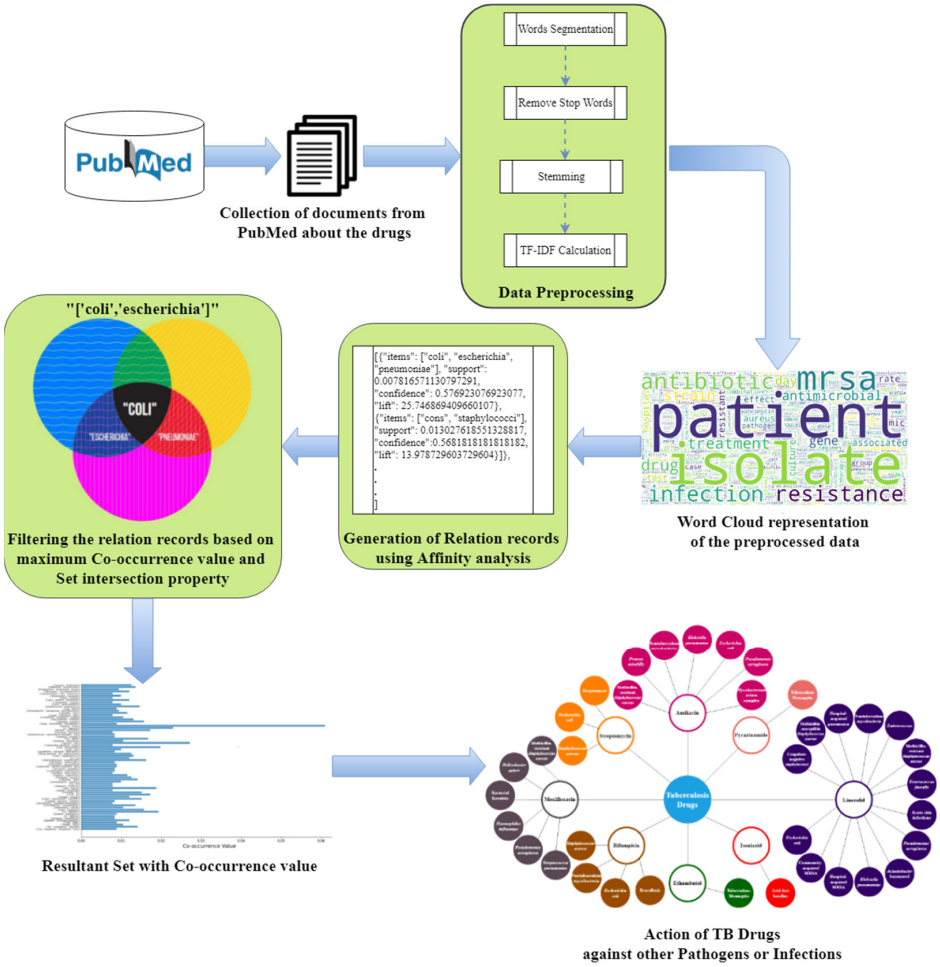


Figure 1. TREASURE framework for the analysis of TB drugs.

3.1. Data Preprocessing

The TREASURE model gathers data from the PubMed database. As the data gathering methods are loosely controlled, the gathered data might contain garbage values, out-of-range values, missing values, etc. This data must be transformed to a useful and an efficient format. It is done by various preprocessing steps. Algorithm 2 shows the TREASURE data preprocessing. The gathered data is first cleaned to handle the noisy and missing data. Then, tokenization breaks the sentences or clauses into separate words. The punctuations among the sentences are removed. In any dataset, there are always some commonly used words such as "the", "for", "should", and "to", which do not contribute to any kind of learning. They are known as stop words. Removing them is an important step in data preprocessing. Inpython, NLTK (Natural Language Toolkit) library has a list of stop words. Along with this, an additional list of stop words, if given as user input, is also removed. The dataset might contain words like affects, affecting, affected, while all these words mean the same thing. These words are either reduced to their root words or word stems that affixes to suffixes or prefixes, and this process is referred as stemming.

The next step in data preprocessing is to extract the important words through tf-idf calculation. Once the words are reduced to their stem the term frequency (tf) and the inverse document frequency (idf) are computed for each word as given in Equations (1) and (2). The tf-idf weight is a statistical measure used to determine the importance of a word (term) to a document (doc) in the dataset [25,26]. Its calculation is given in Equation (3).

$$\mathrm{tf}_{(\mathrm{term,\ doc})} = \frac{\text{frequency of } term \text{ in } doc}{\text{Total number of words in } doc} \quad (1)$$

$$\text{idf}_{(\text{term})} = \ln \frac{\text{Total number of documents}}{\text{Number of documents containing term}} \quad (2)$$

$$\text{tf} - \text{idf}_{(\text{term, doc})} = \text{tf}_{(\text{term, doc})} \times \text{idf}_{(\text{term})} \quad (3)$$

Algorithm 2 TREASURE Data Preprocessing

Input: PubMed abstracts in a csv file
Output: Preprocessed data in a csv file
1. Loop through the entire csv file
   1.1 Perform tokenization on the document
   1.2 Remove punctuations
   1.3 Remove stop words
   1.4 Stem the tokenized words to get the root words
2. Calculate term frequency for the words in document as given in Equation (1)
3. Calculate inverse document frequency as given in Equation (2)
4. Compute tf-idf values for the words as given in Equation (3)
5. Set a minimum threshold value for tf-idf
6. Open a new csv file
   6.1 For row_i write each word of document_i whose tf-idf > threshold

3.2. Generation of Relation Records Using Affinity Analysis

The preprocessed data must be integrated to derive any inference. For example, Amikacin is used in the treatment of non-tuberculous mycobacterial (NTM) disease. This sentence exhibits the relationship between amikacin and non-tuberculous mycobacteria. This kind of co-occurrence relationships can be determined by the affinity analysis technique. Algorithm 3 shows the generation of relation records using affinity analysis. In this technique, we consider each document as a transaction and each word as an item. To determine various connections between items, some formal definitions of measures like support, confidence and lift are needed [27–29].

Support is a simple and yet an important metric in affinity analysis. Its equation is given in Equation (4). The support of (A ∪ B), where A and B are item sets, is given as the ratio of all the transactions that contain all items of (A ∪ B) to the number of transactions in the dataset.

$$Support\ (\text{A} \rightarrow \text{B}) = Support\ (\text{A} \cup \text{B}) = \text{P}\ (\text{A} \cup \text{B}) \quad (4)$$

Confidence denotes the likelihood of certain items to occur together. It is given in Equation (5) and is defined as the proportion of transactions containing item set A that also contain item set B.

$$Confidence\ (\text{A} \rightarrow \text{B}) = \text{P}(\text{A} \mid \text{B}) = \frac{Support\ (A \cup B)}{Support\ (A)} \quad (5)$$

Lift is another important measure in affinity analysis. It is the ratio of probability of A and B occurring together to the product of probabilities of A and B occurring as if there was no association between them Equation (6).

$$Lift\ (\text{A} \rightarrow \text{B}) = \frac{Support\ (A \cup B)}{Support\ (A)\ Support\ (B)} \quad (6)$$

Algorithm 3 TREASURE Relation Records Generation via Affinity Analysis

Input: Preprocessed csv file, minimum number of items in a set as min_length, minimum co-occurrence value as min_support and the minimum conditional property as min_cofidence
Output: A JSON file containing a list of relation records with corresponding confidence, support and lift values
1. Read each item in the file
2. Calculate support for every item as given in Equation (4)
3. Insert every item into a frequent dataset whose support $\geq$ min_support
4. For each item in the frequent dataset calculate confidence and lift values as given in Equations (5) and (6)
5. Insert every rule into a JSON file whose confidence and items count are greater than the corresponding threshold

3.3. Filtering Relation Records Based on Maximum Co-Occurrence Value and Set Intersection Property

The generated relation records contain interrelated words with corresponding support, confidence and lift values. However, not all the relations generated, will have a useful meaning. Therefore, we need to filter these records to extract the primary combination of words. For this purpose, the set intersection property is applied on the records. Among the intersecting sets, the one with the maximum co-occurrence value, i.e., the relation that has frequently occurred, is filtered out and added to the resultant set. This is done to prevent the repetition of same inference and to obtain as many unique inferences as possible. For example, ['capreomycin', 'injectable'], ['capreomycin', 'tuberculosis'] are the intersecting sets. The inference obtained here is that capreomycin is an injectable drug used in the treatment of TB. Among these, ['capreomycin', 'injectable'] has the maximum co-occurrence value and thus, it is added to the resultant set. Algorithm 4 shows the filtration of relation records. By applying this technique, the most essential relations among the records are filtered out and various inferences are obtained.

Algorithm 4 TREASURE Relation Records Filtration based on Maximum Co-occurrence Value and Set Intersection

Input: A list of relation records from the JSON file as D Output: A filtered resultant set S containing frequently occurring word patterns
1. Initialize an empty dictionary ED and an empty set ES
2. For each i in range (length (D))
   - 2.1 For each j in range (i + 1, length (D))
     - 2.1.1 Initialize an empty set IS
     - 2.1.2 Find D[i] $\cap$ D[j] and store the result in IS
     - 2.1.3 If length (IS) > 0 then
       - Store D[i] and D[j] as a key/value pair in ED such that an object capable of holding various items is associated with each key as value
       - Put D[i] and D[j] as individual elements in ES
3. Initialize an empty set S
4. While ES is not empty
   - 4.1 Find an element s with maximum co-occurrence value in ES
   - 4.2 Add the element s to S
   - 4.3 Find all the elements which on set intersection with s is not null from ED
   - 4.4 Remove those elements and the element s from the set ES
5. Display the resultant set S

4. Results

4.1. Data Preprocessing

Around eight drugs are analyzed with this model. For this purpose, abstracts from each document are collected, as they provide the accurate and necessary information about the paper. The PubMed abstracts have a unique ID called PMID. The metapub library in

python gets these IDs as input and extracts their corresponding abstracts. The number of document abstracts collected for each drug from PubMed is given in Table 1.

Table 1. Number of documents collected from PubMed for each drug.

Name of the TB Drug	Number of Document Abstracts
Pyrazinamide	1566
Moxifloxacin	1947
Ethambutol	1841
Isoniazid	1896
Rifampicin	2209
Linezolid	1919
Streptomycin	1954
Amikacin	1909

The NLTK package in python is used to perform tokenization, stemming and to remove stop words. The tf-idf value is calculated for the words obtained and the minimum threshold is set around 0.03 to remove some unnecessary words. This threshold value is set after various trial and errors in the range 0.02 to 0.05. The results are then stored in a csv file such that each row corresponds to an abstract and each row contains the preprocessed words of that abstract. Similarly, eight different csv files are created. Figure 2 shows the word cloud representation of the preprocessed data obtained for eight different TB drugs.

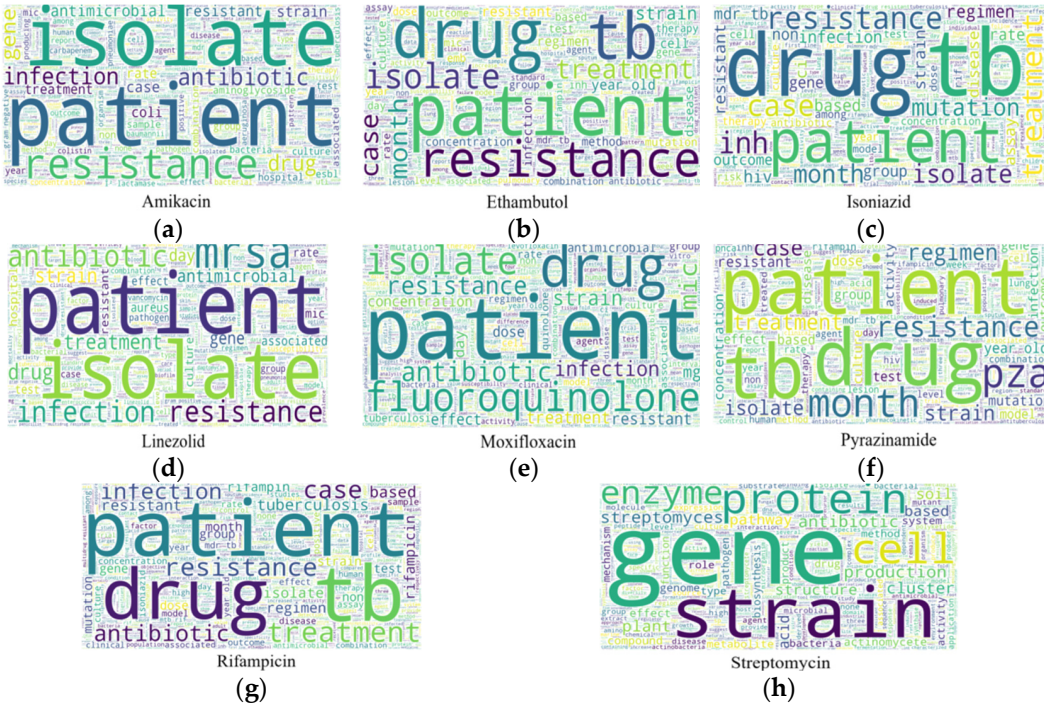


Figure 2. Word clouds for TB drugs: (**a**) Amikacin, (**b**) Ethambutol, (**c**) Isoniazid, (**d**) Linezolid, (**e**) Moxifloxacin, (**f**) Pyrazinamide, (**g**) Rifampicin, (**h**) Streptomycin.

Among the preprocessed data, some words frequently occur than the other. The frequency of the word indicates its importance. Word clouds help to identify such words, as the size of each word in the cloud indicates its frequency of occurrence. Therefore, the preprocessed data are visualized in the form of a word cloud to identify such important words.

4.2. Generation of Relation Records Using Affinity Analysis

This method is used to determine the relationship between the preprocessed data. Each document is considered as a transaction and each word in the document is considered as an item. The apyori library in python is used to perform the affinity analysis. The support, confidence and lift values are computed to determine various connections between the items. The frequently occurring trends among the data are identified.

The minimum co-occurrence value (support) is set around 0.007 after various trials from 0.004 to 0.008 value range, the minimum conditional probability (confidence) is set as 0.5 after various trial and errors in the range 0.4 to 0.6 and the threshold for minimum number of items in the set is kept as 1 after various trials in the range of 1 to 3. The generated relation records are then stored in a JSON file. Therefore, eight JSON files with corresponding relation records are obtained. Tables 2–9 represent the sample item sets obtained for different TB drugs datasets with corresponding support, confidence and lift values.

Table 2. Sample item sets of moxifloxacin dataset.

Item Sets	Support	Confidence	Lift
[“activity”, “bactericidal”]	0.026708	0.626506	5.892789
[“crossover”, “subjects”]	0.010272	0.526316	11.778584
[“aeruginosa”, “p”]	0.014895	0.547170	6.018868
[“genitalium”, “mycoplasma”]	0.010786	0.875	32.143868
[“Gram-negative”, “Gram-positive”]	0.017976	0.538462	14.167360
[“intraocular”, “surgery”]	0.011299	0.564103	15.690110
[“isoniazid”, “pyrazinamide”]	0.018490	0.590164	14.363115
[“methicillin-resistant”, “mrsa”]	0.010786	0.777778	33.651852

Table 3. Sample item sets of amikacin dataset.

Item Sets	Support	Confidence	Lift
[“acinetobacter”, “baumannii”]	0.0230487	0.619718	11.485847
[“aeruginosa”, “pseudomonas”]	0.023573	0.75	10.527574
[“avium”, “clarithromycin”]	0.007333	0.518518	13.747942
[“breakpoints”, “isolates”]	0.009953	0.575758	2.818259
[“cancer”, “patients”]	0.008905	0.586207	3.806357
[“cerebrospinal”, “fluid”]	0.009429	0.9	44.053846
[“adverse”, “events”]	0.008381	0.727273	30.181818
[“coli”, “mirabilis”]	0.007333	0.518518	5.209746

Table 4. Sample item sets of pyrazinamide dataset.

Item Sets	Support	Confidence	Lift
[“abdominal”, “pain”]	0.016603	0.619047	15.892271
[“activity”, “antimycobacterial”]	0.014049	0.611112	5.800001
[“antiretroviral”, “hiv”]	0.012771	0.625	12.083334
[“cerebrospinal”, “fluid”]	0.011494	0.899999	29.987234
[“cough”, “fever”]	0.015964	0.625	15.535714
[“fluoroquinolone”, “mdr-tb”]	0.012132	0.542857	7.143817
[“genes”, “mutations”]	0.019157	0.535714	6.554129
[“hepatic”, “liver”]	0.009578	0.576923	12.906593

Table 5. Sample item sets of ethambutol dataset.

Item Sets	Support	Confidence	Lift
[“antiretroviral”, “hiv”]	0.013036	0.705882	12.866627
[“biopsy”, “patient”]	0.024986	0.511112	4.376537
[“cell”, “wall”]	0.025529	0.72307	12.558345
[“fluid”, “meningitis”]	0.008691	0.551724	18.809706
[“hiv”, “virus”]	0.012493	0.511112	14.167360
[“imaging”, “resonance”]	0.009234	0.679999	27.81955
[“katg”, “mutations”]	0.017381	0.680851	11.292313
[“lymph”, “nodes”]	0.016295	0.967741	32.992831

Table 6. Sample item sets of linezolid dataset.

Item Sets	Support	Confidence	Lift
[“abscessus”, “mycobacterium”]	0.011985	0.696969	17.369933
[“anti-tb”, “drugs”]	0.011464	0.758621	10.109674
[“aureus”, “methicillin-sensitive”]	0.008337	0.761904	6.356935
[“bedaquiline”, “drug-resistant”]	0.013548	0.553191	11.538852
[“ca-mrsa”, “ha-mrsa”]	0.007816	0.882352	48.378151
[“chromosome”, “isolates”]	0.011985	0.766666	4.008811
[“drugs”, “second-line”]	0.013548	0.684211	9.118055
[“faecium”, “isolates”]	0.019801	0.5	2.614441

Table 7. Sample item sets of streptomycin dataset.

Item Sets	Support	Confidence	Lift
[“acetate”, “extract”]	0.008701	0.548387	17.566367
[“chromosome”, “genome”]	0.007676	0.576923	7.320179
[“ethyl”, “extract”]	0.008188	0.533333	17.084153
[“ribosomally”, “synthesized”]	0.007676	0.882352	42.051649
[“crystal”, “structure”]	0.013306	0.722223	11.959511
[“coli”, “e”]	0.012794	0.641025	20.202646
[“biosynthesis”, “inactivation”]	0.006653	0.565217	5.549923
[“aureus”, “methicillin-resistant”]	0.009211	0.692307	32.994371

Table 8. Sample item sets of rifampicin dataset.

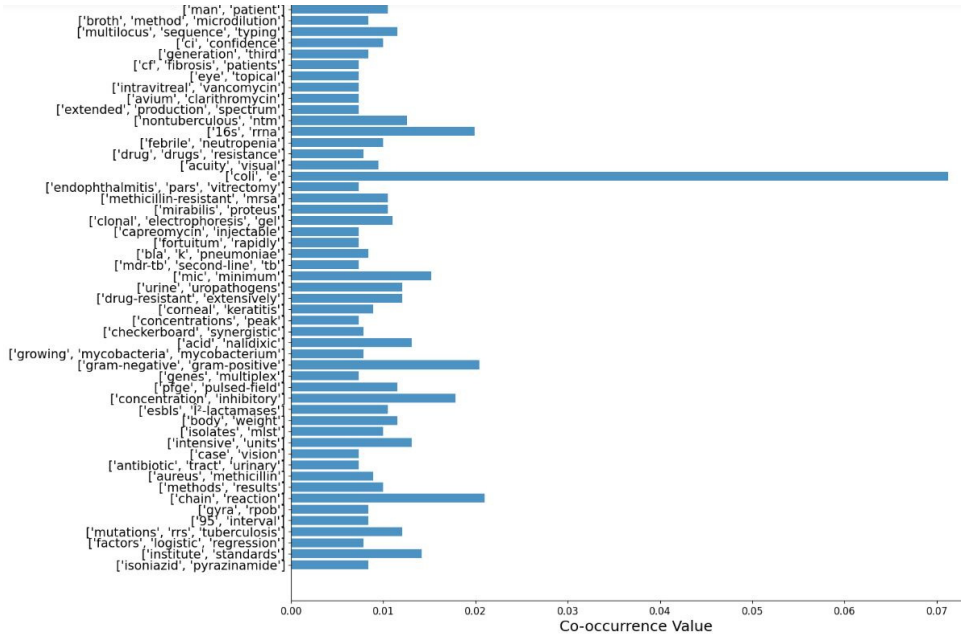
Item Sets	Support	Confidence	Lift
[“activity”, “bactericidal”]	0.009053	0.645161	11.681649
[“antiretroviral”, “hiv”]	0.008601	0.542857	12.491369
[“bacteria”, “Gram-negative”]	0.009053	0.540541	9.950451
[“central”, “nervous”]	0.006791	0.833333	40.018115
[“chest”, “x-ray”]	0.009053	0.606061	21.947342
[“co-infection”, “hiv”]	0.007243	0.64	14.726667
[“codon”, “isolates”]	0.006791	0.75	7.363333
[“injury”, “liver”]	0.01177	0.604651	19.935439

Table 9. Sample item sets of isoniazid dataset.

Item sets	Support	Confidence	Lift
[“activity”, “bactericidal”]	0.010021	0.5	8.697247
[“care”, “hiv”]	0.018459	0.5	6.236842
[“chest”, “x-ray”]	0.008438	0.516129	17.168081
[“drug-induced”, “injury”]	0.013713	0.541667	18.017543
[“fluoroquinolone”, “resistance”]	0.009493	0.545454	3.262403
[“hepatic”, “liver”]	0.010548	0.606061	13.207941
[“interferon-gamma”, “release”]	0.009493	0.9	33.458823
[“nontuberculous”, “ntm”]	0.008438	0.666666	48.615384

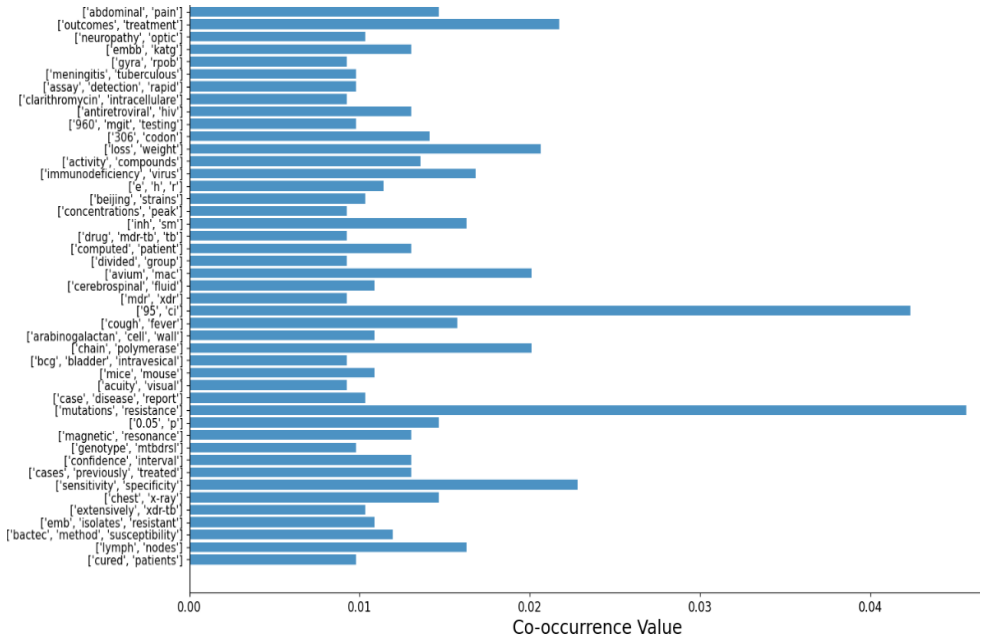
4.3. Filtering Relation Records Based on Maximum Co-Occurrence Value and Set Intersection Property

The generated JSON files consist of the relation records. These records are filtered to determine the most weighted relationships. The python sets are used for this purpose. The set intersection is applied among the records, to group similar relations and to obtain unique inferences at the end. Then, the relation with maximum co-occurrence value from each group of records is extracted and added to the resultant set. Therefore, eight different filtered resultant sets, each corresponding to a drug, are obtained through this method. Figure 3 shows the resultant sets obtained for eight different TB drugs with co-occurrence value corresponding to each item of each set. The records present in the resultant set provide some important information about the drug. For example, in Figure 3c, [‘genes’, ‘inha’, ‘katg’] is one of the elements of the resultant set obtained for the isoniazid dataset. The following inference is obtained for this record. *M.tb* poses a great challenge at the scientific level as it acquires gene mutations, which develop resistance to the drugs and treatment forms. The mutations at codon 315 (amino acid position) of the katG gene are associated with high resistance to isoniazid. Therefore, isoniazid is ineffective in the treatment of *M.tb* with this mutation [30]. Whereas the inhA gene mutations are associated with low-level resistance to isoniazid and thus high doses of the drug can be used for *M.tb* treatment with this mutation [31]. Similarly, various inferences can be obtained from the resultant sets of 8 different TB drugs. Thus, the filtered resultant sets provide the most essential and unique inferences of the corresponding dataset.

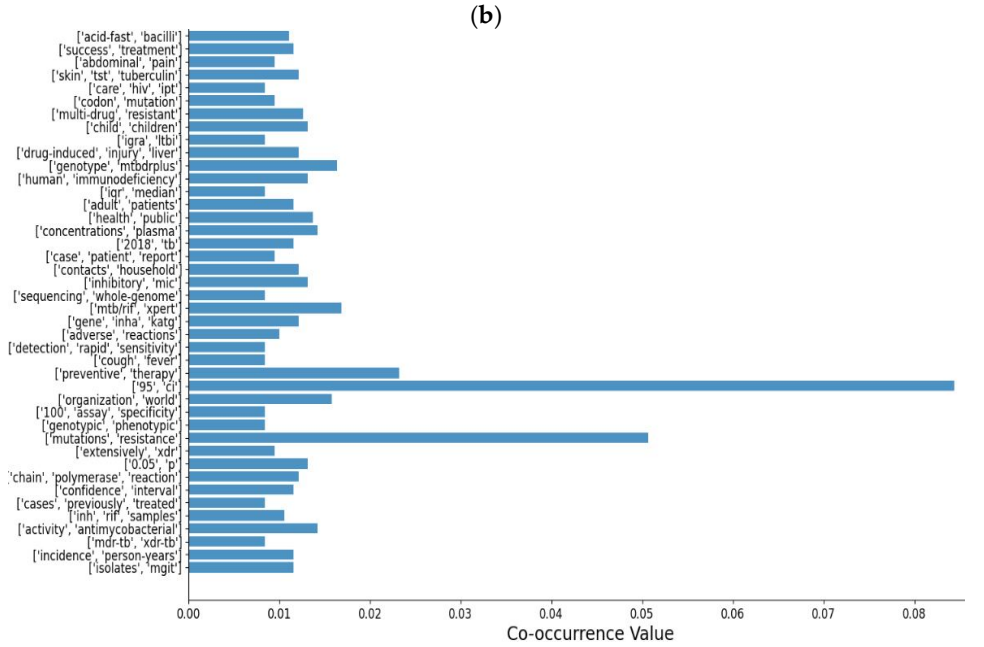


(**a**)

Figure 3. *Cont.*

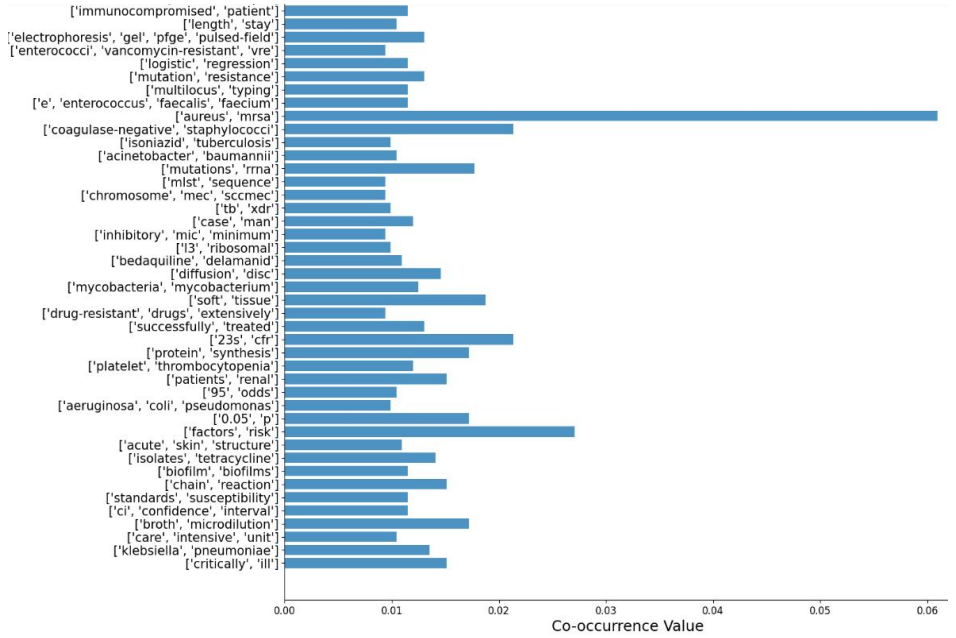


(**b**)

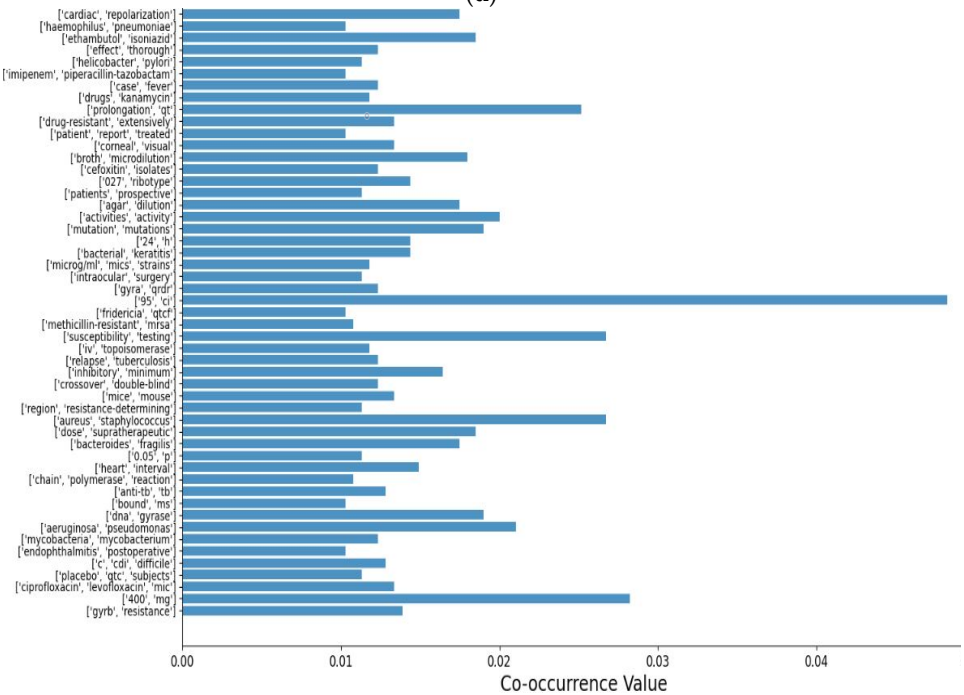


(**c**)

Figure 3. *Cont*.

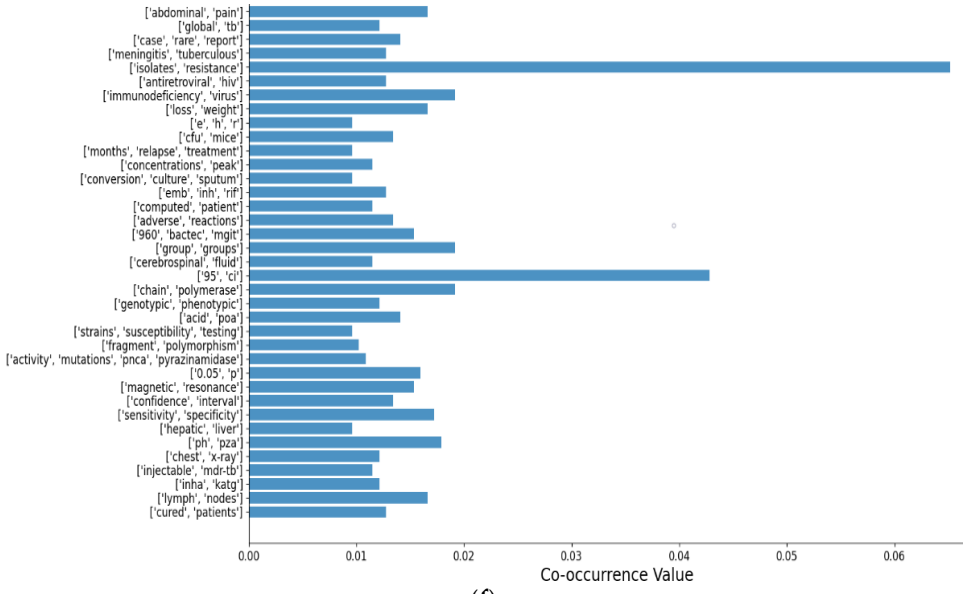


(**d**)

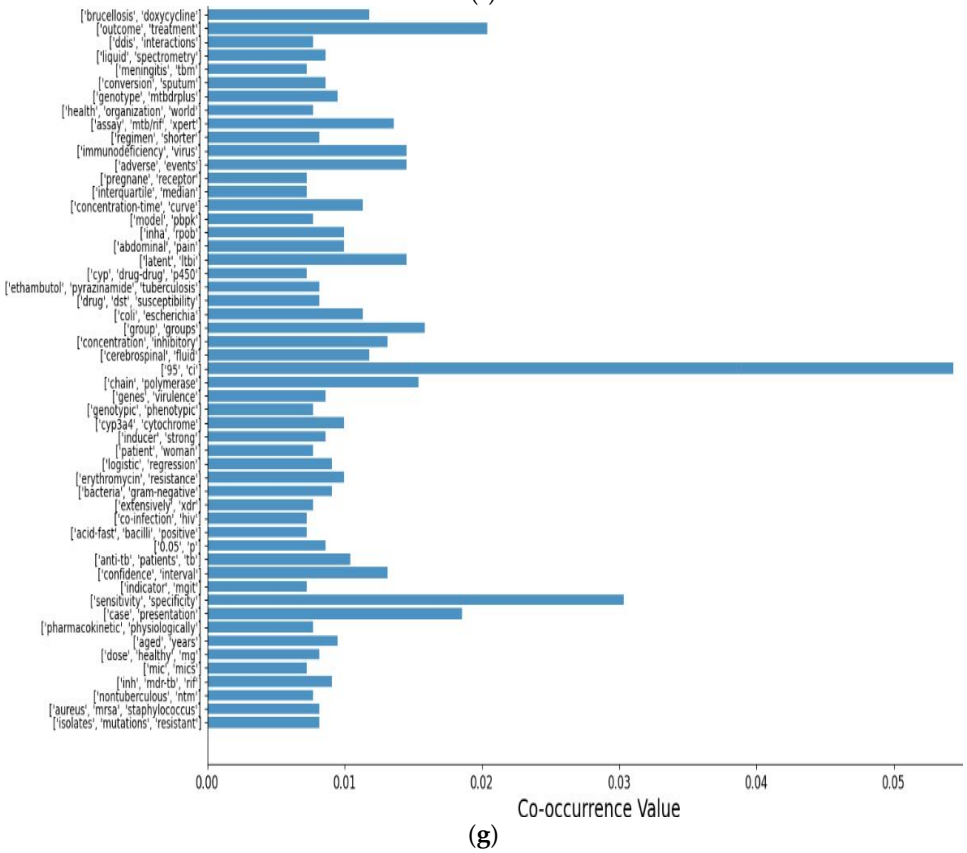


(**e**)

Figure 3. *Cont*.

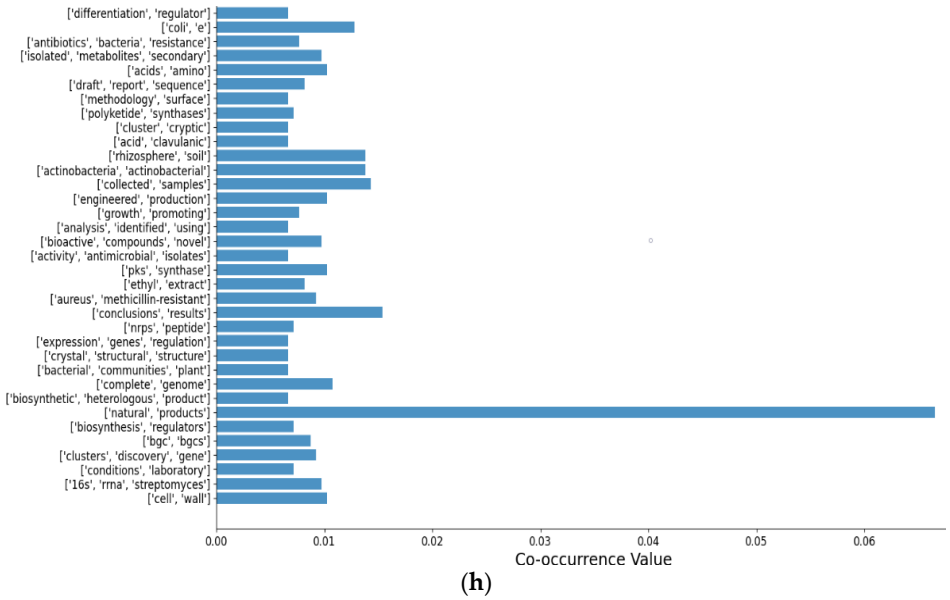


(**f**)



(**g**)

Figure 3. *Cont*.



(**h**)

Figure 3. Resultant sets with co-occurrence value for: (**a**) amikacin; (**b**) ethambutol; (**c**) isoniazid; (**d**) linezolid; (**e**) moxifloxacin; (**f**) pyrazinamide; (**g**) rifampicin; (**h**) streptomycin.

Since the relation records are generated using affinity analysis, they do not contain much noise. The resultant set is obtained through set intersection property and hence, there will not be any repetition of records. The obtained resultant set contains various inferences. These inferences can help the medical experts in their research and can also lead to some discoveries. Suppose if this method is applied on different disease datasets, we may extract some common features and relationships between the diseases. In this paper, the resultant sets of eight different TB drugs are analyzed and compared, to discover such inferences. Though these drugs are developed to work against the pathogens causing TB, there are chances that they might work against other infections. Extracting such inferences may help in the development of new treatments for some diseases.

5. Discussion

The resultant sets are analyzed and an important inference is extracted about the drugs. It determines the pathogens that these drugs are active against, in addition to TB as given in Figure 4. The antibiotic drugs developed for a particular disease work against the pathogen causing that disease. It inhibits the growth of the microbial targets without harming the host. Each class of antibiotics has a unique mechanism of action. They can be the inhibitors of cell wall synthesis, inhibitors of protein synthesis, inhibitors of cell membrane function, inhibitors of nucleic acid synthesis or inhibitors of other metabolic processes [32]. It depends on the nature of their structure and their affinity to the target. A mechanism of action of a drug can work against many pathogens. Therefore, that drug can be used in the treatment of diseases caused by such pathogens. For example, moxifloxacin is also effective against *Helicobacter Pylori*, a bacterium which causes ulcer and might progress to stomach cancer [33,34]. Thus in addition to TB, moxifloxacin can also be used in the treatment of these two diseases.

A Gram-positive bacterium named Methicillin-Resistant *Staphylococcus aureus* (MRSA) causes various infections such as skin infections, pneumonia and sepsis [35]. The drugs, moxifloxacin, amikacin and linezolid are effective against MRSA in addition to TB and thereby can be used in the treatment of the mentioned infections. Similarly the TB drugs can be effective against various other pathogens or infections. Thereby they can be used in the treatment of various diseases caused by such pathogens. This kind of TB drug–pathogen relationship is extracted through our framework. The results are compared with some important information sources such as doctors and medical research papers, and they

closely associate with the inferences obtained from these sources [36–39]. This inference might help medical researchers in various drug analyses, in finding treatment for various other bacterial diseases, etc. Researchers who carry out in vitro testing can also benefit from this study by extracting this inference from about 1000 papers, which would have otherwise been extremely cumbersome under usual settings.

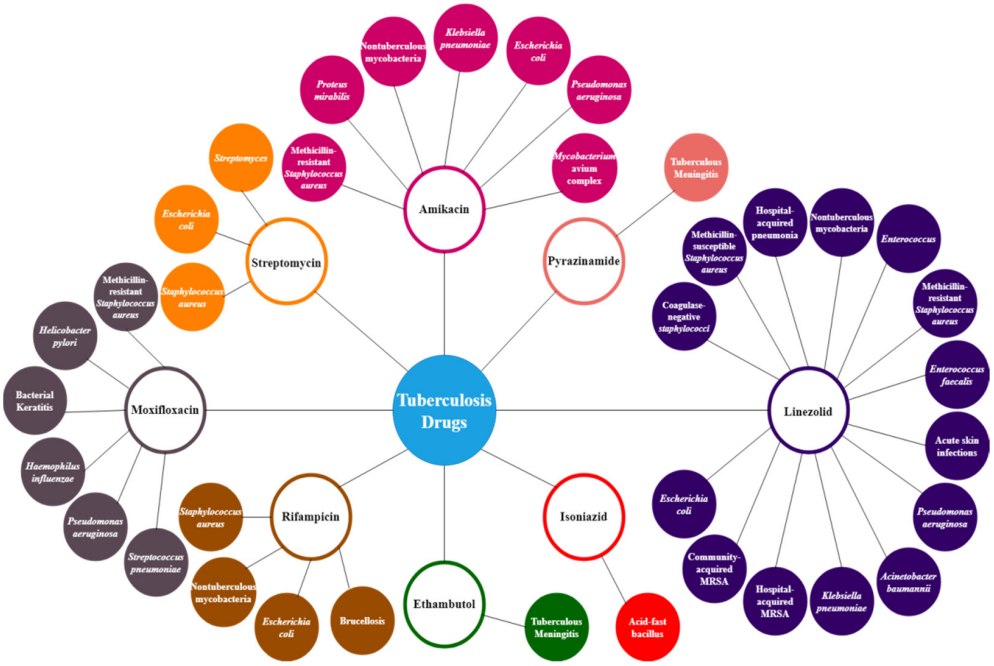


Figure 4. Action of TB drugs against other pathogens or infections.

Another result is also obtained through this framework is depicted in Figure 5. The resultant sets of the mentioned TB drugs in this figure, contained either antiretroviral or immunodeficiency virus in them. This inference states that some of the TB drugs are also used in antiretroviral therapy to treat Human Immunodeficiency Virus (HIV). This inference shows the inter-relationships between TB and HIV. People diagnosed with HIV have a high chance of getting infected with TB pathogens as HIV weakens the immune system [40,41]. Therefore, the body cannot fight TB, thereby these pathogens can quickly progress into TB disease. TB has become the leading cause of death among the people affected by HIV. If a person is affected by TB, it is important for them to know about their HIV status.

Similarly, various inferences can be extracted through this framework. It is evident that this way of data extraction has identified many different and important results. This tool when applied to other drugs, diseases or plants datasets can draw out various important inferences and thereby can help the medical researchers to speed up their background research work. This framework can be considered as a complementing tool for doctors and medical research experts in their drug discovery studies, drug interaction studies and bacteria analysis, and can pave a way in discovering treatment for some incurable diseases.

In order to predict the accuracy of the proposed framework, it is compared with the existing topic modelling techniques such as latent Dirichlet allocation (LDA), latent Dirichlet allocation with affinity propagation (LDA with AP) and latent semantic analysis (LSA). The performance of our framework is evaluated using the following standard methodologies [42,43].

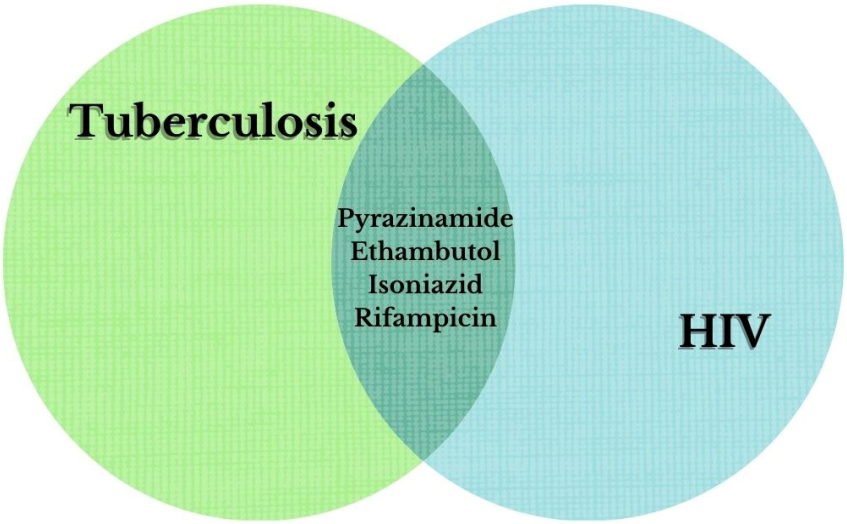


Figure 5. A Venn diagram summarization to depict the inter-relationships between the TB and HIV treatments.

Precision is the measure of how much information returned by the system is correct. It is given in Equation (7). It is the ratio of the number of correctly predicted observations to the total predicted observations in the resultant set.

$$Precision\% = \frac{number\ of\ correct\ answers\ given\ by\ the\ system}{number\ of\ answers\ given\ by\ the\ system} \times 100 \quad (7)$$

Recall, as given in Equation (8), the measure of relevant information extracted by the system. It is the ratio of correctly predicted observations to all the observations in the class.

$$\text{Recall\%} = \frac{number\ of\ correct\ answers\ given\ by\ the\ system}{total\ number\ of\ possible\ correct\ answers\ in\ text} \times 100 \quad (8)$$

F-Measure is the harmonic mean of precision and recall and is computed as given in Equation (9). It balances both the precision and recall with a single score.

$$\text{F} - \text{Measure\%} = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100 \quad (9)$$

The above mentioned standard methodologies are calculated for various algorithms. The graphical comparisons of these measures for four different drug datasets, moxifloxacin, linezolid, streptomycin and rifampicin are depicted in Figure 6a–c. From Figure 6c, we can observe that the TREASURE model gives high accuracy (71.85%) when compared to LDA, LDA with AP and LSA algorithms. The benchmark words for comparing the results of each algorithm were collected from renowned information sources such as WHO, Healthline, MEDLINE and NIH. Hence, it is evident that this way of data extraction has provided appropriate results in the identification of various TB drug–pathogen relationships.

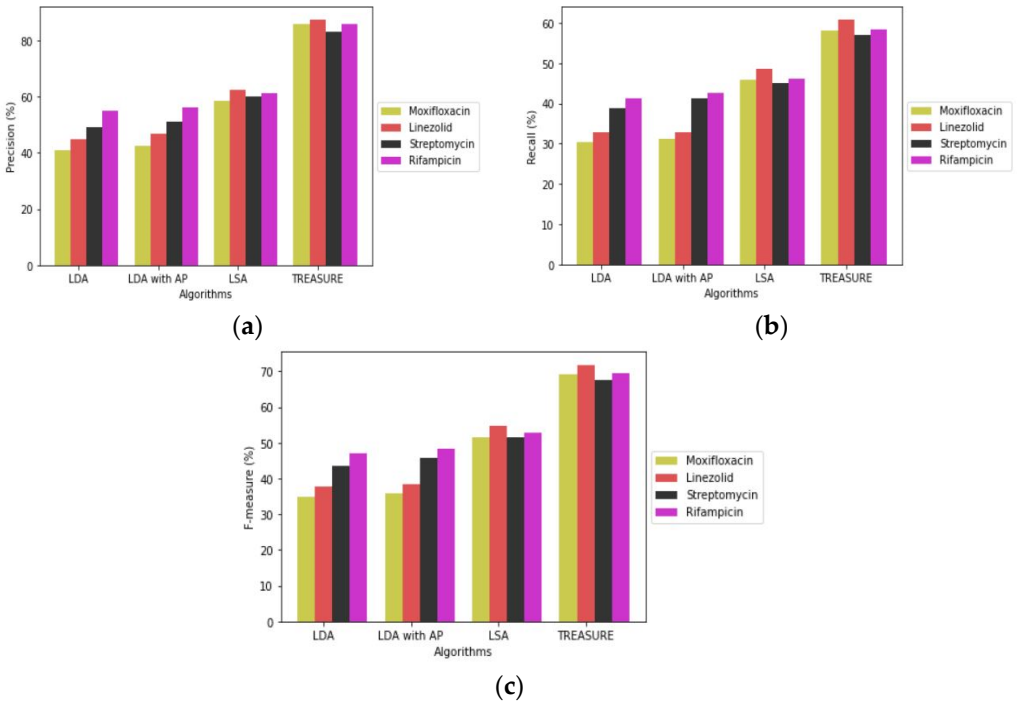


Figure 6. Comparison of various algorithms for the following measures: (**a**) precision; (**b**) recall; (**c**) F-measure.

6. Conclusions

Tuberculosis is a potentially serious infectious disease caused by *Mycobacterium tuberculosis* that usually attacks the lungs. This paper is a novel attempt to propose a framework named TREASURE to analyze various TB drugs from PubMed literature and identify the action of these drugs against other pathogens. Lack of effective analysis tools to discover different drug–pathogen relationships necessitated the proposal of TREASURE model. We analyzed eight different TB drugs namely pyrazinamide, moxifloxacin, ethambutol, isoniazid, rifampicin, linezolid, streptomycin and amikacin, and found out various other pathogens or infections that these drugs are effective against, in addition to TB. We generated relation records from the datasets using affinity analysis and filtered these records using maximum co-occurrence value and set intersection property to obtain the results. This method provides inferences based on various drug–pathogen relationships which can help the medical experts to speed up their background research work and thereby saves time and manpower. In future, it can also be used to find remedies for some incurable diseases. In this application, we use only the text datasets. As a future research, it is intended to combine this model with image processing and analyze its performance.

Author Contributions: Conceptualization, P.S. and N.S.S.; methodology, P.S. and N.S.S.; validation, V.S. and Y.L.; formal analysis, V.S.; investigation, Y.L.; resources, Y.L.; data curation, P.S. and N.S.S.; writing, P.S. and N.S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-02207, Development of ICT system utilizing interactive emotional device for providing intact welfare care service).

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Not Applicable.

Conflicts of Interest: The authors declare that they do not have any conflict of interest. This research does not involve any human or animal participation. All authors have checked and agreed with the submission.

References

1. Gilpin, C.; Korobitsyn, A.; Migliori, G.B.; Raviglione, M.C.; Weyer, K. The World Health Organization standards for tuberculosis care and management. *Eur. Respir. J.* **2018**, *51*, 1800098. [CrossRef] [PubMed]
2. Mazumdar, S.; Satyanarayana, S.; Pai, M. Self-reported tuberculosis in India: Evidence from NFHS-4. *BMJ Glob. Health* **2019**, *4*, e001371. [CrossRef]
3. Motschall, E.; Falck-Ytter, Y. Searching the MEDLINE Literature Database through PubMed: A Short Guide. *Oncol. Res. Treat.* **2005**, *28*, 517–522. [CrossRef] [PubMed]
4. Koch, A.; Mizrahi, V. Mycobacterium tuberculosis. *Trends Microbiol.* **2018**, *26*, 555–556. [CrossRef]
5. Sahbazian, B.; Weis, S.E. Treatment of Active Tuberculosis: Challenges and Prospects. *Clin. Chest Med.* **2005**, *26*, 273–282. [CrossRef]
6. Khan, F.A.; Minion, J.; Pai, M.; Royce, S.; Burman, W.; Harries, A.D.; Menzies, D. Treatment of Active Tuberculosis in HIV-Coinfected Patients: A Systematic Review and Meta-Analysis. *Clin. Infect. Dis.* **2010**, *50*, 1288–1299. [CrossRef]
7. Shi, R.; Itagaki, N.; Sugawara, I. Overview of anti-tuberculosis (TB) drugs and their resistance mechanisms. *Mini-Rev. Med. Chem.* **2007**, *7*, 1177–1185. [CrossRef]
8. Kolyva, A.S.; Karakousis, P.C. Old and New TB Drugs: Mechanisms of Action and Resistance. In *Understanding Tuberculosis—New Approaches to Fighting Against Drug Resistance*; Books on Demand: Norderstedt, Germany, 2012. [CrossRef]
9. Pantziarka, P.; Verbaanderd, C.; Huys, I.; Bouche, G.; Meheus, L. Repurposing drugs in oncology: From candidate selection to clinical adoption. *Semin. Cancer Biol.* **2021**, *68*, 186–191. [CrossRef] [PubMed]
10. Zhu, Y.; Jung, W.; Wang, F.; Che, C. Drug repurposing against Parkinson's disease by text mining the scientific literature. *Libr. Hi Tech* **2020**, *38*, 741–750. [CrossRef]
11. Jin, X.; Wu, Y. Study on Main Drugs and Drug Combinations of Patient-Controlled Analgesia Based on Text Mining. *Pain Res. Manag.* **2020**, *2020*, 1–7. [CrossRef]
12. Naseem, U.; Khushi, M.; Khan, S.K.; Shaukat, K.; Moni, M.A. A comparative analysis of active learning for biomedical text mining. *Appl. Syst. Innov.* **2021**, *4*, 23. [CrossRef]
13. Zhou, J.; Fu, B.-Q. The research on gene-disease association based on text-mining of PubMed. *BMC Bioinform.* **2018**, *19*, 1–8. [CrossRef]
14. Rani, J.; Ramachandran, S. Pubmed. mineR: An R package with text-mining algorithms to analyse PubMed abstracts. *J. Biosci.* **2015**, *40*, 671–682. [CrossRef] [PubMed]
15. Vazquez, M.; Krallinger, M.; Leitner, F.; Valencia, A. Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications. *Mol. Inform.* **2011**, *30*, 506–519. [CrossRef]
16. Wang, H.; Ding, Y.; Tang, J.; Dong, X.; He, B.; Qiu, J.; Wild, D.J. Finding Complex Biological Relationships in Recent PubMed Articles Using Bio-LDA. *PLoS ONE* **2011**, *6*, e17243. [CrossRef] [PubMed]
17. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimed. Tools Appl.* **2019**, *78*, 15169–15211. [CrossRef]
18. Guan, R.; Wen, X.; Liang, Y.; Xu, N.; He, B.; Feng, X. Trends in Alzheimer's Disease Research Based upon Machine Learning Analysis of PubMed Abstracts. *Int. J. Biol. Sci.* **2019**, *15*, 2065–2074. [CrossRef] [PubMed]
19. Vanteru, B.C.; Shaik, J.S.; Yeasin, M. Semantically linking and browsing PubMed abstracts with gene ontology. *BMC Genom.* **2008**, *9*, S10–S11. [CrossRef] [PubMed]
20. Yadav, C.; Sharan, A. A New LSA and Entropy-Based Approach for Automatic Text Document Summarization. *Int. J. Semant. Web Inf. Syst.* **2018**, *14*, 1–32. [CrossRef]
21. Kaur, M.; Sapra, R. Classification of patents by using the text mining approach based on PCA and logistics. *Int. J. Eng. Adv. Technol.* **2013**, *2*, 711–714.
22. Lakshmi, K.; Krishna, M.; Kumar, S. Utilization of Data Mining Techniques for Prediction and Diagnosis of Tuberculosis Disease Survivability. *Int. J. Mod. Educ. Comput. Sci.* **2013**, *5*, 8–17. [CrossRef]
23. Wang, X.; Wang, H.; Xie, J. Genes and regulatory networks involved in persistence of Mycobacterium tuberculosis. *Sci. China Life Sci.* **2011**, *54*, 300–310. [CrossRef]
24. Asha, T.; Natarajan, S.; Murthy, K.N.B. Association-rule-based tuberculosis disease diagnosis. *Second Int. Conf. Digit. Image Process.* **2010**, *7546*, 75462. [CrossRef]
25. Qaiser, S.; Ali, R. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *Int. J. Comput. Appl.* **2018**, *181*, 25–29. [CrossRef]
26. Jing, L.-P.; Huang, H.-K.; Shi, H.-B. Improved feature selection approach TFIDF in text mining. In Proceedings of the International Conference on Machine Learning and Cybernetics, Beijing, China, 4–5 November 2002; Volume 2, pp. 944–946. [CrossRef]
27. Karthiyayini, R.; Balasubramanian, R. Affinity Analysis and Association Rule Mining using Apriori Algorithm in Market Basket Analysis. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2016**, *6*, 241–246.

28. Prajapati, D.J.; Garg, S.; Chauhan, N. Interesting association rule mining with consistent and inconsistent rule detection from big sales data in distributed environment. *Futur. Comput. Inform. J.* **2017**, *2*, 19–30. [CrossRef]
29. Sanida, T.; Varlamis, I. Application of Affinity Analysis Techniques on Diagnosis and Prescription Data. In Proceedings of the 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS), Thessaloniki, Greece, 22–24 June 2017; pp. 403–408.
30. Lempens, P.; Meehan, C.; Vandelannoote, K.; Fissette, K.; De Rijk, P.; Van Deun, A.; Rigouts, L.; De Jong, B.C. Isoniazid resistance levels of Mycobacterium tuberculosis can largely be predicted by high-confidence resistance-conferring mutations. *Sci. Rep.* **2018**, *8*, 1–9. [CrossRef]
31. Bollela, V.R.; Namburete, E.I.; Feliciano, C.S.; Macheque, D.; Harrison, L.H.; Caminero, J.A. Detection of katG and inhA mutations to guide isoniazid and ethionamide use for drug-resistant tuberculosis. *Int. J. Tuberc. Lung Dis.* **2016**, *20*, 1099–1104. [CrossRef]
32. Kaufman, G. Antibiotics: Mode of action and mechanisms of resistance. *Nurs. Stand.* **2011**, *25*, 49–55. [CrossRef]
33. Blaser, M.J. Helicobacter pylori: Its Role in Disease. *Clin. Infect. Dis.* **1992**, *15*, 386–393. [CrossRef]
34. Graham, D.Y. History of Helicobacter pylori, duodenal ulcer, gastric ulcer and gastric cancer. *World J. Gastroenterol.* **2014**, *20*, 5191–5204. [CrossRef]
35. Barcenas, C.G.; Fuller, T.J.; Elms, J.; Cohen, R.; White, M.G. Staphylococcal sepsis in patients on chronic hemodialysis regimens: Intravenous treatment with vancomycin given once weekly. *Arch. Intern. Med.* **1976**, *136*, 1131–1134. [CrossRef]
36. Callegan, M.C.; Ramirez, R.; Kane, S.T.; Cochran, D.C.; Jensen, H. Antibacterial activity of the fourth-generation fluoroquinolones gatifloxacin and moxifloxacin against ocular pathogens. *Adv. Ther.* **2003**, *20*, 246–252. [CrossRef] [PubMed]
37. Montgomery, A.B.; Rhomberg, P.; Abuan, T.; Walters, K.A.; Flamm, R.K. Potentiation Effects of Amikacin and Fosfomycin against Selected Amikacin-Nonsusceptible Gram-Negative Respiratory Tract Pathogens. *Antimicrob. Agents Chemother.* **2014**, *58*, 3714–3719. [CrossRef] [PubMed]
38. Noskin, G.A.; Siddiqui, F.; Stosor, V.; Hacek, D.; Peterson, L.R. In Vitro Activities of Linezolid against Important Gram-Positive Bacterial Pathogens Including Vancomycin-Resistant Enterococci. *Antimicrob. Agents Chemother.* **1999**, *43*, 2059–2062. [CrossRef] [PubMed]
39. Caroline, P.M.; Linezolid, J.B. A review of its use in the management of serious gram-positive infections. *Drugs* **2003**, *63*, 2126.
40. Toossi, Z.; Mayanja-Kizza, H.; Hirsch, C.S.; Edmonds, K.L.; Spahlinger, T.; Hom, D.L.; Aung, H.; Mugyenyi, P.; Ellner, J.; Whalen, C.W. Impact of tuberculosis (TB) on HIV-1 activity in dually infected patients. *Clin. Exp. Immunol.* **2001**, *123*, 233–238. [CrossRef]
41. McShane, H. Co-infection with HIV and TB: Double trouble. *Int. J. STD AIDS* **2005**, *16*, 95–101. [CrossRef]
42. Nahm, U.Y.; Mooney, R.J. Text mining with information extraction. In Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, Stanford, CA, USA, 25–27 March 2002; pp. 60–67.
43. Yang, Y. An Evaluation of Statistical Approaches to Text Categorization. *Inf. Retr.* **1999**, *1*, 69–90. [CrossRef]